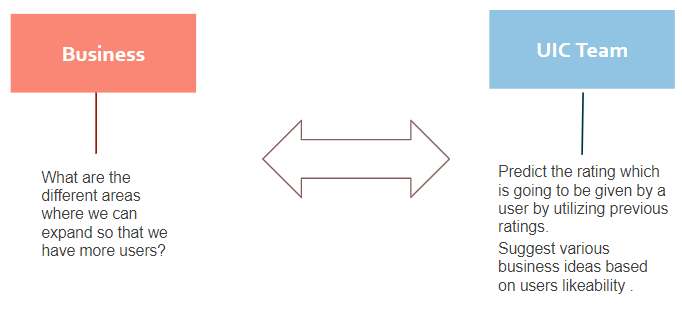


**Predictive Analytics using Collaborative filtering on Yelp Data**

**Problem Setting:** The goal of this project is to compare different types of businesses and create an automated recommendation system for users. This involves predicting rating values of business that users have not visited before based on their historical rating records. Based on this rating, we recommend top business to users which can be explored based on their likability. We built an Alternative Least Square recommendation system. The performance of our models is mainly measured by Root Mean Square Error (RMSE). We used Google Colab as a platform and visualized the results on Tableau.



**Data Description:** We downloaded the dataset from kaggle and It can be accessed through this link <https://www.kaggle.com/yelp-dataset/yelp-dataset>. The data comes in a zip file. The file contains 6 JSON data files and our models mainly use the following:

* **reviews.json:** contains 7 million reviews from 1.6 million users across 190 thousand businesses in 10 metropolitan areas. Each record is a rating and a full text review that the user gives to a business on a certain date time
* **business.json:** contains business profile including location data, attributes, and categories
* **user.json:** contains user profile including the user’s friend mapping and all the metadata associated with the user e.g. reviews count, age of account, number of cool votes received,etc

**Techniques:**

* **Data Pre-processing:** Relevance of each column was indicated based on requirement of ALS algorithm from all the data files
* **Data Cleaning:** Leverage the data pre-processing step to filter the relevant columns
* **Data Integration:** Convert all the varchars into integers to merge into single database
* **Master Database:** Create metrics based on these data sources to test hypotheses

**Modelling Technique:**

Collaborative filtering is commonly used for recommender systems and is used as the technique to predict the future rating based on current user ratings. Predictions (filtering) are made about the interests of a user by collecting preferences or taste information from many users (collaborating). Spark MLlib library provides the Collaborative filtering implementations using Alternating Least Squares. The implementation in MLlib has the following parameters:

**⦁ numBlocks:** The number of blocks the users and items will be partitioned into in order to parallelize computation (defaults to 10)

**⦁ rank:** It is the number of latent factors in the model

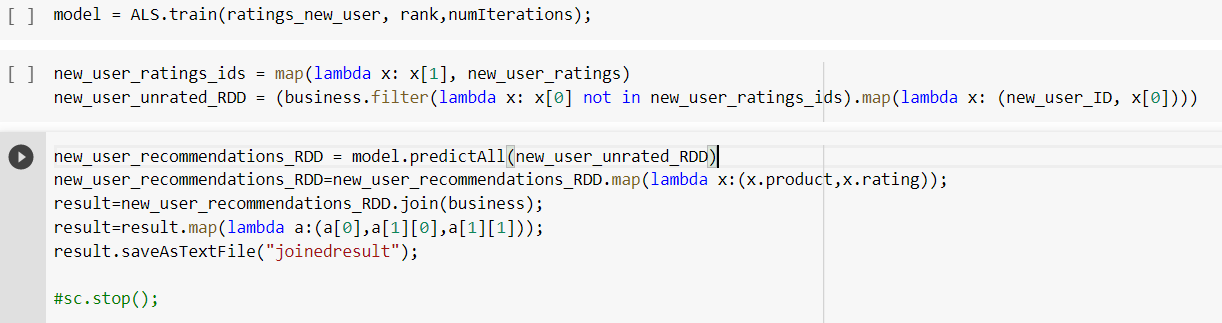
**⦁ iterations:** It is the number of iterations to run

**⦁ lambda:** It specifies the regularization parameter in ALS

**⦁ implicitPrefs:** It specifies whether to use the explicit or implicit feedback ALS variant

**⦁ alpha:** It is applicable to implicit feedback variant of ALS to govern the baseline confidence

**ALS Model:**



**Results:** For results, we usedTableau as our visualization tool to plot the final results. From the below given visual we can infer that “New Garden Restaurant” has the highest user ratings followed by “Hilton Garden” based on the popularity rankings which is decided by analyzing the historical data of users likeability.

